International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 9, ISSN No. 2455-2143, Pages 59-63
Published Online July – August 2016 in IJEAST (http://www.ijeast.com)

HANDLING COLD START PROBLEM IN RECOMMENDER SYSTEMS BY CLUSTERING DEMOGRAPHIC ATTRIBUTE

Sugandha Gupta
Department of Computer Science & Engg.
Thapar University, Patiala, Punjab, India

Shefali Arora
Department of Computer Science & Engg.
Thapar University, Patiala, Punjab, India

Abstract— **Recommender engines have become immensely important in recent years because a large number of people depend on internet to browse options out of a vast set of choices. Different websites implement recommender systems using different techniques such as content-based filtering, collaborative filtering or hybrid filtering. Recommender systems face various challenges like scalability problem, cold start problem and sparsity issues. Cold start problem arises when no sufficient information is available for the user who has recently logged into the system and no proper recommendations can be made. In this paper, a novel approach is introduced which deals with cold start in recommenders on the basis of demographic attributes (Age, Gender, Occupation) and similarity among users by using k-means and hierarchical clustering. The clustered datasets are classified further using efficient techniques. Weka and MySQL are used as the tools to deal with this problem. Thus it is an effective approach to provide recommendations when no user ratings are available.**

Keywords— **collaborative filtering, cold start, recommender system, k-means clustering, Weka**

I. INTRODUCTION

Recommender engines are those systems which predict preferences of a user based on his previous ratings of items. These systems are finding their place in E-commerce websites and online retail stores as rage of internet is growing day by day and choices are getting huge. The major advantages of having a recommender system are customer retention, information retrieval, personalization and many more. Also recommender systems can be used in domains like books, restaurants, movies etc. to recommend products to users. Some successful websites which have established the use of recommender systems are Movielens, Amazon, ebay, Linkedln and Facebook.

Recommender systems compare similarities between user profiles and predict the user ratings for an item that they had not yet thought of. The ratings could be taken from the user in an explicit or implicit manner.

Recommender systems are categorized into the following three basic categories:

- Collaborative Filtering Recommender Systems (also referred to as social filtering): The information is filtered by comparing tastes of a user with other users [1]. It works on the notion that people with similar tastes or preferences in the past are likely to agree in the future again. Thus Collaborative filtering is based on correlation among a given set of users.

- Content-based Recommender System: These systems recommend items which are similar to the items which the user rated or used before [2] i.e. the predictions are on the basis of the content of the previously preferred item. The content of the item is symbolized by terms that are used to represent the user profiles, which are further analyzed to make recommendations.

- Hybrid Recommender Systems: These use integration of two techniques explained above, which could be a better method of recommendation in various cases.

Cold start problem is a special type of sparsity problem when a user or item has no ratings. When a new user or item is added to the system, with no history of purchase or ratings present in the system. Thus it is not possible to generate recommendations for them. Cold start problem is of two kinds:

i) New User cold start problem: It arises when there are no ratings for the user who has arrived into the system. New item cold start problem: It occurs when an item has entered the system but not rated as of yet.

In this paper, we focus on new user cold start problem i.e. when a new user enters the system and there are no recommendations available [3]. This is done by applying k-means and hierarchal clustering to the demographic attributes available in the user dataset i.e. Age, Gender and Occupation. In the second phase, the clustered instances are classified using efficient techniques J-48(C4.5 algorithm)[4] and Naïve Bayes classifier[5] using Weka and evaluation metrics like Precision, Recall, F1 ,MAE and correctly classified instances are recorded. The final phase is to generate recommendations using PhpMyAdmin and MySQL for the new user on the basis of the cluster he is placed in. Thus the proposed model:

- Handles new user cold start problem
- Involves similarity metrics in calculation
- Does not require complicated calculations

The paper is divided into these sections: Section II gives the details of previous work done on cold start problem. Section III gives the description of experiments applied to the user dataset and Section IV gives the analysis of these results. Section V is a conclusion of the work.

II. RELATED WORK

Many researchers have done a plenty of work in the area of recommender systems. In their work, Adomavicius and Tuzhilin has given an overview of collaborative , content and hybrid recommender systems along with the limitations of current recommendation techniques and their possible extensions[6]. The basic concepts of collaborative filtering, its drawbacks and a clustering algorithm for huge datasets is proposed by Sarwar et al.[7]. Sanchez et al. analyzed Pearson correlation metric and cosine metric together with the less common mean squared difference in order to discover their advantages and disadvantages[8]. Also, Schein et al. have given a new evaluation metric known as the CROC curve and explained numerous components of testing strategies empirically to obtain a good performance of recommender systems[9].

A number of approaches have been proposed to deal with the cold start problem. In [10] data from collaborative and content based filtering is combined by incorporating users, items and contents of items. A probabilistic model is applied to influence the recommendations generated.

The work done by Ling Yanxiang et al.[11] explains how to address cold start problem using character capture and clustering methods[12]. In [13] five classification strategies are used and the results are merged using exponential type of ordered weighting operator.(OWA) . Various approaches are considered, like using user rating history, use of demographic features of the user and information about item attributes. If final output in the form of a label for item x and corresponding user y is 1, the item is recommended else it is not.
OWA operator has been introduced as:

$$\text{F: } R^n \rightarrow R \text{ and is given by} \\ \text{OWA}\,(a_1, a_2, \dots\dots, a_n) = \textstyle\sum_{i=1}^{n} w_i b_i \qquad (1)$$

where b_i is the ith largest element among a_i's. The sum equals 1 and weights are positive. $(\sum_{i=1}^{n} w_i = 1)$.

Lam et al. [14] gave a combination of collaborative and content based filtering is done by using probabilistic models to address new user cold start problem.

Authors in[15] make use of association rules to expand user profiles and use non redundant rule set to avoid cold start. Zhou et al. [16] used functional matrix factorization to enable a user to query the recommender system. A decision tree is created where each node is the interview question. This interview phase could be alleviated to deal with cold start and improve performance of the system.

Park and Chu [17] used feature based regression models help to avail all the required information about users and items to deal with cold start problem in recommenders.

III. EXPERIMENTAL ANALYSIS

A. K-means and Hierarchial clustering techniques-

Considering N data points which are to be partitioned into k disjoint subsets S_j with N_j data points. The sum of square is minimized by using the formula:

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2 , \qquad (2)$$

Here x_n is a vector which indicates nth data point and μ_j represents the geometric centroid of data points in the obtained subsets. K-means algorithm works on discrete variables rather than continuous to achieve a minimum J over the clustered points. The algorithm works in the following steps:

- Data points are assigned to random k sets
- Every data point is given to a cluster whose centroid is nearest to this point.
- This is repeated till there is no change in the assignments of clusters.

On the other hand, hierarchal clustering is done in the following steps:

- Each item is assigned to a cluster. Thus for N points there are N clusters.
- The closest pair of clusters are merged into a single cluster.
- Similarities are calculated between each of the new and old clusters.
- Repeat the steps until they form a single cluster of size N.

In this paper, K-means and hierarchical clustering are applied to our dataset which contains information about users and their demographic attributes. Further, Naïve Bayes and J-48 classification algorithms are used to find how many clusters are classified correctly.

J48 works on C4.5 algorithm which builds decision trees from training data. It chose a data attribute that effectively splits the sample into subsets from one class or the other. On the other hand, Naïve Bayes classifier is a conditional probabilistic model classifier which assigns class labels to instances and uses Bayes theorem to classify instances [5].

B. Dataset description and Experimental setup-

Movielens dataset ml-1M is chosen for the purpose of our experiments. It contains around 1000000 anonymous ratings of approximately 4000 movies given by 6040 users. These ratings go back to as far as the year 2000.It describes a 5-star rating activity by the users from movie recommender service, Movielens (http://movielens.org), a movie recommendation service. For our convenience, we take a subset of 999 users to carry out the clustering process in WEKA. Other features of the experimental setup are:

- The number of clusters is set to 3 in both cases.
- The dataset is split into training and testing sets. 66% of the data is taken for the purpose of training and the remaining is tested to show the final results.
- The demographic attributes involved are Gender, Age and Occupation of the users.

C. Evaluation Metrics-

The metrics used to evaluate the classification of clustered instances are MAE (Mean absolute error), RMSE (Root Mean Squared Error), Precision and Recall.
Given test set T and user-item pair (u, i), MAE give the average absolute difference for actual rating and predicted rating is:

$$MAE = \sqrt{1/T \sum_{(u,i)\in T} |\hat{r}} \qquad (3)$$

Whereas RMSE is calculated as:

$$RMSE = \sqrt{1/|T| \sum_{(u,i)\in T} (\hat{r}_{ui} - r_{ui})^2} \qquad (4)$$

For precision and recall, the following measures are often used: True Positives (TP) i.e. the number of instances in a class that actually belong to it; True Negatives (TN) i.e. the number of instances that do not belong to a class and are not classified thus; False Positive (FP) i.e. the number of instances classified into a particular class although they do not belong to it and False Negatives (FN) i.e. instances classified into a particular class but do not fall in it. As per the following measures, the accuracy of the system can be defined as the number of classified instances o the total instances. Precision is defined as the measure of number of errors made while classifying instances into a particular class. Recall measures the capability of not leaving out the instances that need to be classified in a particular class.

$$Precision = \frac{TP}{TP + FP} \qquad (5)$$

$$Recall = \frac{TP}{TP + FN} \qquad (6)$$

D. Results-

The user dataset consisting of 999 users is clustered using K-means and Hierarchical clustering methods in Weka. For each of the results shown in Figure 1 and 2, the clustered instances are classified using Naïve Bayes and J48 algorithm. The final results are evaluated using metrics like Precision, Recall, MAE and RMSE. It is found that J48 classification work best when the user dataset is clustered using the hierarchical clustering method in Weka. Figures 1 and 2 show the results of J48 classification on hierarchical and simple k-means clusters. and The comparison between the evaluation metrics for Naïve Bayes classification and J48 classification are shown in Table 1 and 2.

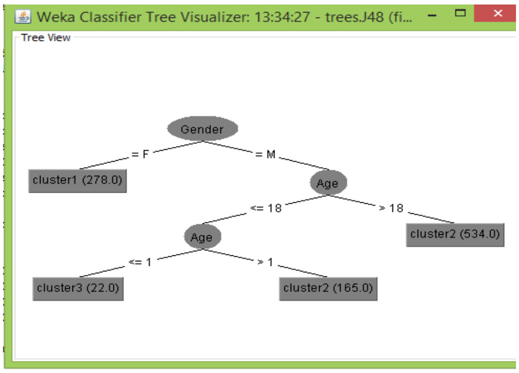

Fig. 1. J48 classification on hierarchical clusters

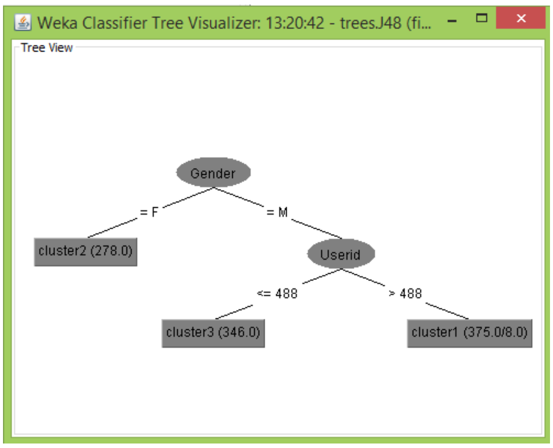

Fig. 2. J48 classification on simple k-means clusters

Table -1 Evaluation metrics for Hierarchical clustering

	Precision	Recall	MAE	RMSE
Naïve Bayes	1.0	1.0	0.0057	0.0391
J-48	1.0	1.0	0	0

Table -2 Evaluation metrics for K-means clustering

	Precision	Recall	MAE	RMSE
Naïve Bayes	0.992	0.986	0.0315	0.0933
J-48	0.981	1.0	0.0093	0.0682

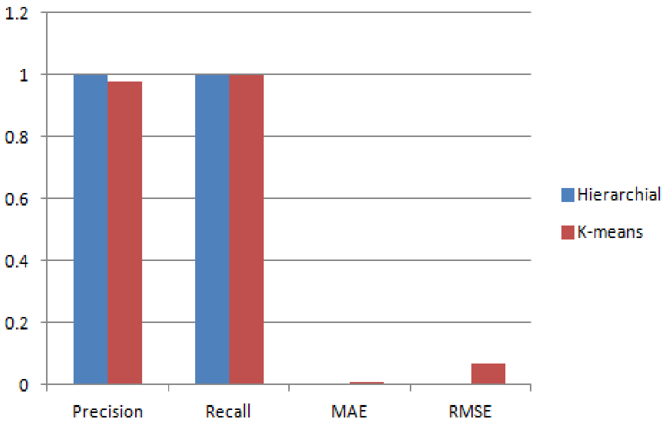

Fig. 3. A comparison of clustering methods using J48 algorithm

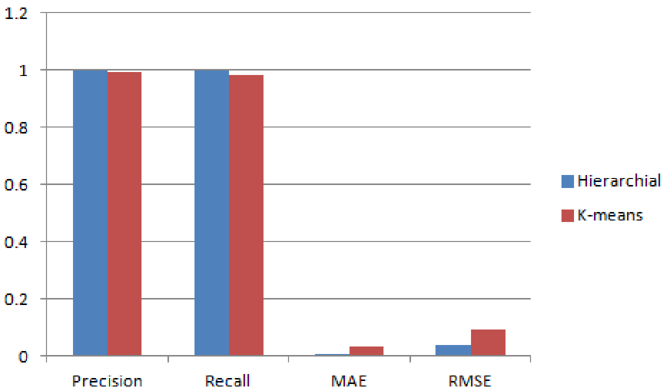

Fig. 4. A comparison of clustering methods using Naïve Bayes algorithm

Figures 3 and 4 shows that Hierarchical clustering works best on the user dataset comprising 999 users when the clustered instances are classified using J48 algorithm. This is because the accuracy of classification is 100% and thus good quality of recommendations is obtained by the new user who enters the system.

E. Recommendations-

Once the users have been divided into clusters, recommendations are generated using phpMyAdmin and MySQL. The new user who enters the system thus gets recommendations on the basis of querying or aggregate analysis of ratings in MySQL.

IV. CONCLUSION

In this paper, cold start problem has been addressed by applying clustering in WEKA. By applying evaluation metrics like RMSE, MAE, Precision and Recall, it is found that hierarchical clustering works best for the user dataset of Movielens. J48 classification of these clustered instances gives the most precise and error free results. Thus a new user can be accurately given recommendations on the basis of his cluster when he enters the system.

As a future scope, this technique could be applied to items involved in the recommendation process. This technique only solves user based cold start problem. More efficient clustering approaches could be designed to address this issue.

V. REFERENCE

[1] P. Andrei-Cristian, “Implementation Of A Recommender System using Collaborative Filtering,” STUDIA UNIV. BABES_{BOLYAI,
INFORMATICA, Volume LV,Number 4,2010,pp. 70-84

[2] K. Iwahama,Y. Hijikata and S. Nishida, “Content-Based Filtering For Music Data,”in Proceedings of the 2004

International Symposium on Applications and Internet Workshops(SAINTW'04),2000,pp. 1-8.

[3] B. Lika, K. Kolomvatsos and S. Hadjiefthymiades, "Facing the cold start problem in recommender systems," Expert Systems with Applications, Volume 41, Part 2, March 2014, pp.2065-2073.

[4] S.B. Kotsiantis, "Supervised Machine Learning:A review of classification techniques,"in Informatica, 31, pp. 249-268.

[5] H. Zhang, "The optimality of Naïve Bayes",in FLAIRS,2004.

[6] G. Adomavicius and A. Tuzhilin , " Toward The Next Generation Of Recommender Systems: A Survey Of The State-Of-The-Art and Possible Extensions,"in IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 6, 2005,pp.734-749.

[7] B. Sarwar,J. Karypis,J.Konstan and J. Riedl, "Item Based Collaborative Filtering Algorithms,"in Proceedings of the 10th International Conference On World Wide Web,ser WWW'01,ACM,New York,NY,USA,2001,pp. 285-295.

[8] J.L. Sanchez, F. Serradilla, E. Martinez and J. Bobadilla, " Choice Of Metrics Used In Collaborative Filtering and Their Impact On Recommender Systems," in Digital Ecosystems and Technologies ,2nd IEEE International Conference ,2008,pp. 432–436.

[9] A.I. Schein, A. Popescul, L.H. Ungar, and D.M. Pennock, "Methods and Metrics for Cold-Start Recommendations,"In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '02,New York, NY, USA,ACM ,2002,pp. 253–260.

[10] A. Popescul, L.H.Ungar, D.M. Pennock and S. Lawrence, "Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments,"In Proceedings of the 17th conference on uncertainty in artificial intelligence,2001.

[11] L.Yanxiang , G. Deke , C. Fei and C.Honghui, "User-based Clustering With Top-N Recommendation On Cold-Start Problem," In Intelligent System Design and Engineering Applications (ISDEA), Third International Conference ,2013,pp. 1585–1589.

[12] M.K.K Devi, R.T. Samy, S.V. Kumar and P. Venkatesh, " Probabilistic Neural Network Approach To Alleviate Sparsity and Cold Start Problems In Collaborative Recommender Systems," In Computational Intelligence and Computing Research (ICCIC), IEEE International Conference ,2010,pp. 1–4.

[13] H. Jindal and S.K. Singh , " A Hybrid Recommendation System for cold start problem using Online Commercial Dataset,"in International Journal Of Computer Engineering And Applications,Vol VII, No 1,July 2014.

[14] X.N Lam, T. Vu,T.D. Le and A.D Duong, " Addressing cold-start problem in recommendation systems,"In Proceedings of the second international conference on ubiquitous information management and, communication,2008.

[15] G. Shaw, Y. Xu and S. Geva, " Using association rules to solve the cold-start problem in recommender systems,"In Lecture notes in computer science,2010.

[16] K. Zhou, S.H. Yang and H. Zha , " Functional matrix factorizations for cold-start recommendation," In Proceedings of the 34th international ACM SIGIR,2011

[17] S.T. Park and W. Chu, " Pairwise preference regression for cold-start recommendation," In Proceedings of the third ACM conference on recommender systems,2009.